	CL 7B	CL 13B	CL 34B	GPT 3.5	GPT 4	Human Written	Own *vs* LLMs	Own *vs* LLMs \ F	Own *vs* Human
Content Adequacy									
human	2.95	3.05	2.78	3.99	4.54	2.93	-	-	-
CL 7B	1.48	1.36	1.55	0.44	-0.15	1.53	*** (S)	*** (L)	(N)
CL 13B	1.32	1.08	1.29	0.32	-0.21	1.40	* (S)	*** (M)	(N)
CL 34B	1.47	1.09	1.45	0.31	-0.11	1.30	*** (M)	*** (L)	(N)
GPT 3.5	1.17	0.85	0.98	0.33	0.11	0.75	** (S)	*** (M)	** (S)
GPT 4	-0.09	-0.35	-0.28	-0.05	-0.13	-0.46	(N)	(N)	(N)
Conciseness									
human	3.95	4.49	4.43	4.92	4.20	4.91	-	-	-
CL 7B	0.22	-0.36	-0.48	-0.69	-0.20	-0.67	*** (M)	*** (M)	*** (M)
CL 13B	-0.39	-0.89	-0.80	-1.37	-0.78	-1.19	(N)	(N)	(N)
CL 34B	0.20	-0.45	-0.46	-0.77	-0.09	-0.74	(N)	(N)	(S)
GPT 3.5	-0.79	-1.37	-1.45	-1.46	-0.65	-2.04	*** (S)	* (N)	*** (M)
GPT 4	-0.58	-0.80	-1.06	-0.16	-0.08	-0.80	*** (M)	*** (M)	*** (L)
Fluency&Understandability									
human	3.89	4.06	3.71	4.65	4.68	3.92	-	-	-
CL 7B	0.50	0.44	0.66	-0.14	-0.27	0.63	** (S)	*** (M)	(N)
CL 13B	0.57	0.32	0.67	-0.30	-0.35	0.34	(N)	*** (S)	(N)
CL 34B	0.34	0.09	0.61	-0.34	-0.42	0.20	*** (M)	*** (L)	* (S)
GPT 3.5	0.51	0.31	0.37	0.05	0.12	0.00	* (S)	** (S)	(N)
GPT 4	0.25	0.06	0.19	0.31	0.32	0.01	(N)	(N)	(S)

Adjusted p-values: * <0.05, ** <0.01, *** <0.001. Cliff delta: N=Negligible, S=Small, M=Medium, L=Large

Table 1: **(Python)** Code Summarization: Average of differences between the LLM judgments and the ground truth (*i.e.,* the median of the score given by the three human raters). Last three columns report adj. p-value and effect size when comparing the judgments each LLM gave to summaries it generated against those it gave when judging summaries (i) generated by all other LLMs, (ii) generated by all other LLMs but those belonging to the same family, and (iii) written by humans.